Program 3 Report

Shane Stevens

For my report, I decided to train my model using the 2014 data, and then test each year individually to see how data drift affected the accuracy of my model. I found, not surprisingly, that the data drifted more and more the farther the year was from 2014. The accuracy of the 2014 data was 0.9873333333333333 for logarithmic regression, and 0.9036666666666666 for random forest. Below is the data for the years 2015-2021

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

While in general, the data drift got worse with each year, there were a few exceptions. 2016 is slightly more accurate than 2015, although the difference is fairly insignificant. Both years have around .90 accuracy, which tells me that there wasn’t much data drift between 2014, 2015, and 2016, and likely little to know drift between 2015 and 2016.

2019 also shows a massive decrease in accuracy, which as we discussed in class, is due to the Android API being changed in 2019. Interestingly enough, for 2019 onwards, the random forest model is more accurate than the logarithmic regression model, unlike for the previous years where logarithmic regression was more accurate than random forest. If I had to guess why this is, I think that random forest might be slightly more resistant to data drift since it uses several different trees to determine classes, instead of just using one logarithmic regression model to determine classes.

In 2021, The accuracy increases compared to 2020 for both models. This implies that the data drifted back to be a little more similar to 2014’s data, which I find interesting. My finding shows the importance of making a training model with data drift in mind. Below is the confusion matrices for my data.

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated